TEACH FOR AMERICA PREDICTIVE ANALYTICS PROJECT

GROUP 3

Ana Manuela Bustamante Vela
Carolina Guerrero Requejo
Nihaal Karkera
Pauline Punio
Ana Uribarri
Tachin Ho

Table of Contents

1

PROBLEM BACKGROUND

Origin of “Teach For America”

Metzger's Challenge: As TFA (Teach for America) faces declining application and offer acceptance rates, Metzger and his team are working to leverage data-driven insights to optimize the recruiting process.

Issue with Current Model: A model intended to predict the likelihood of applicants accepting or declining offers caused misinterpretations. For example, a recruiter scheduled unnecessary meetings with all applicants after a model incorrectly predicted an applicant's withdrawal risk. This misstep reduced efficiency in recruiting efforts.

Core Business Problem

Declining Applications & Offer Acceptance Rates: TFA needs to address the decrease in both the volume of applications and the acceptance rate of offers.

Inefficient Recruiting Efforts: Over-scheduling meetings with applicants based on inaccurate model predictions reduces recruiter efficiency.

Data Misinterpretation by Non-Technical Staff: Non-technical staff struggle to interpret model predictions, leading to suboptimal decisions and wasted resources.

Objective

Develop Predictive Models

Build models using TFA's 2015-2016 applicant data to predict which applicants are at risk of withdrawing during the selection process.

Optimize Recruiter Time

Use these predictions to prioritize recruiting activities, ensuring that recruiters focus on high-risk applicants to reduce withdrawals and improve overall efficiency.

2

DATA CLEAN UP

Clean Up

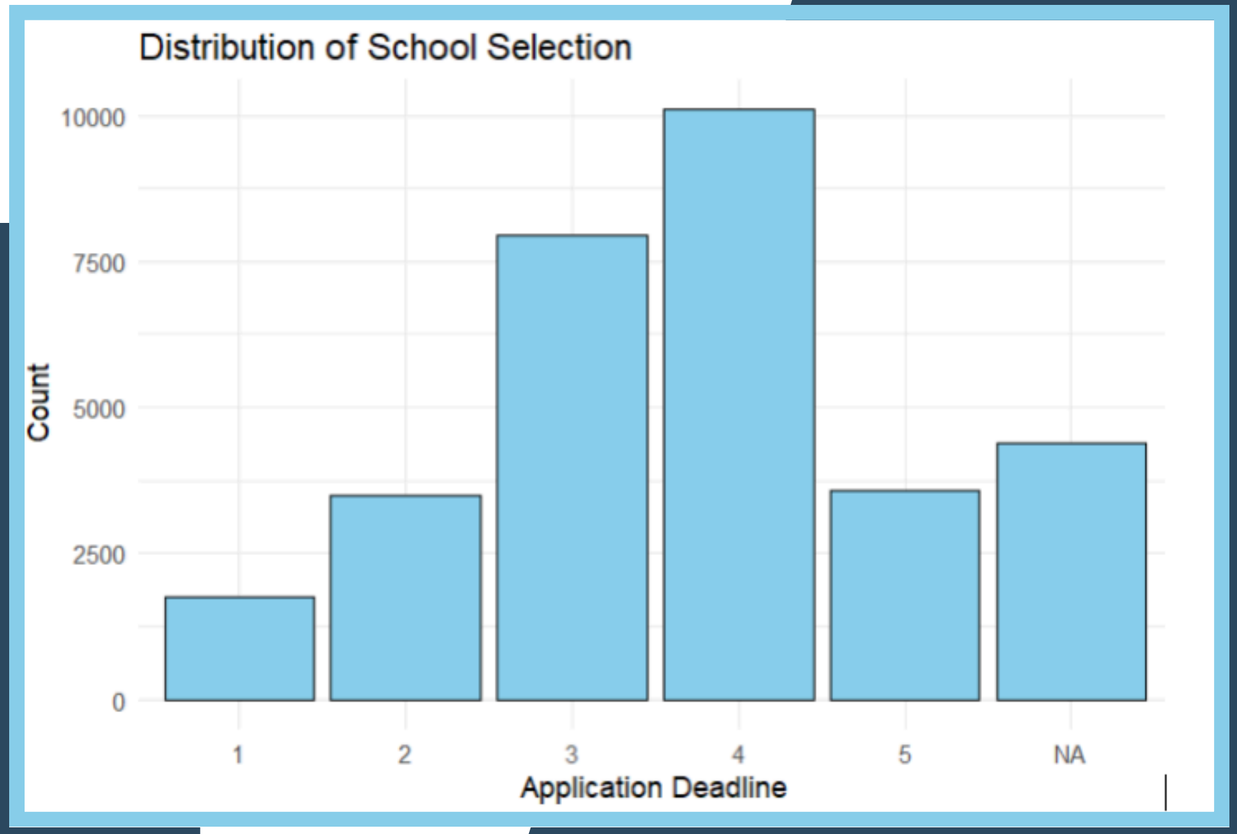
1. Removed variables:

personid, appyear, and schoolsel_chr

2. Schoolsel variable:

replace all missing (NA) values with 0

all across the models for NB, Decision Tree, ANN, and SVM.



3

NAIVE BAYES

Evaluating different Naive Bayes models to determine the best feature subset for predicting applicant behavior.

MODEL 1

Reference / prediction	0	1
0	81	142
1	1158	4868

Sensitivity: 0.06538
Specificity: 0. 97166

Summary of Results:

- Specificity: High, indicating good prediction of completions.
- Sensitivity: Low, meaning the model struggled to predict withdrawals.
- Overall accuracy was **79.2%**

MODEL 2

Reference / prediction	0	1
0	30	38
1	1209	4972

Sensitivity: 0.024213
Specificity: 0.992415

Summary of Results:

- Variables: stem, schoolsel, major1group, etc.
- Accuracy improved slightly to **80.04%**.
- Specificity remained high, but Sensitivity was even lower.

MODEL 3:

Reference / prediction	0	1
0	22	26
1	1217	4984

Sensitivity: 0.017756
Specificity: 0. 994810

Summary of Results:

- Variables: stem, schoolsel, major1group, etc., reducing more features from the previous model.
- Accuracy was similar to Model 2 (**80.11%**), while Specificity stayed high and Sensitivity remained low.

Refining Naive Bayes models through cross-validation and parameter tuning to enhance predictive stability.

MODEL 4

Reference / prediction	0	1
0	80	110
1	1159	4900

Sensitivity: 0.06457
Specificity: 0.97804

Summary of Results:

- This model used trainControl with 10-fold cross-validation.
- Accuracy was **79.69%**, comparable to earlier models.
- Specificity remained high, while Sensitivity stayed consistently low.

MODEL 5

Reference / prediction	0	1
0	81	112
1	1158	489

Sensitivity: 0.06538
Specificity: 0. 97764

Summary of Results:

- Hyperparameter tuning was applied using a grid search to optimize performance.
- There was no significant improvement in accuracy (**79.68%**), though Specificity remained high.

CONCLUSIONS

- Strong performance in predicting completions (high specificity).
- Sensitivity remained a challenge across all models.
- Fourth model with cross-validation showed the best overall performance.
- Further improvement: experiment with advanced models to enhance sensitivity.

4

DECISION TREE

Decision Tree Model

Objective: Use a Decision Tree model to predict whether an applicant successfully completed the admissions process (completedadm).

Initial Approach

- Initially used all available variables, resulting in a tree size of 690.
- Such a large tree indicated this model was overly complex and not prcatical.

After several trials, we decided to use a smaller, **more selective set of features for better accuracy and interpretability.**

Selected Features

- **Academic Features**: fgpa (categorical GPA), schoolsel (school selectivity), major1group, major2group.
- **Application Features**: appdeadline (application deadline), attendedevent (event attendance), submitted (submission timing).
- **Essay Feature**s: essayUW (unique words), essentiment (essay sentiment).
- **Target Variable**: completedadm.

Data Split

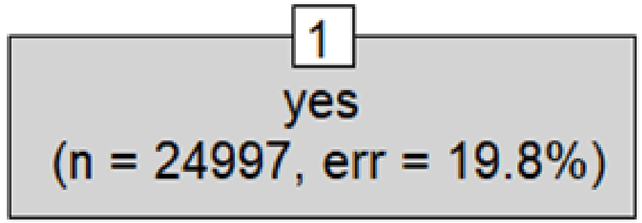
- Dataset split into 8**0% training** and **20% testing sets**.
- The **class distribution was preserved** between training and testing sets, with approximately **19.8%** of instances being no (did not complete) and **80.2%** being yes (completed).

This ensured that both sets represented the original data distribution accurately.

Initial Decision Tree Model - Model 1

Overview of the Model:

- **Algorithm**: C5.0 Decision Tree
- **Data**: Trained with 24,997 samples and 9 predictors.
- **Model Complexity**: Resulted in a tree size of only 1, which was overly simplistic.



Reference / prediction	0	1
0	0	1239
1	0	5010

Inability to properly differentiate between applicants who completed versus withdrew from the admissions process.

- **Accuracy:** 0.8017
- **Sensitivity:** NA (ability to identify applicants likely to withdraw)
- **Specificity:** 0.8017 (ability to identify applicants likely to complete)

Improvement of the Decision Tree Model - Model 2 // Final Model

Initial Challenge: Class imbalance affected model performance.
Boosting Attempt: Applied boosting with 100 trials but abandoned due to **ineffectiveness**
Solution: Cost Matrix Implementation.

Introduced a cost matrix to penalize incorrect predictions..

Penalty Structure:

- Specifically, the cost of predicting "no" as "yes" was set higher than the cost of predicting "yes" as "no."

Costs assigned:

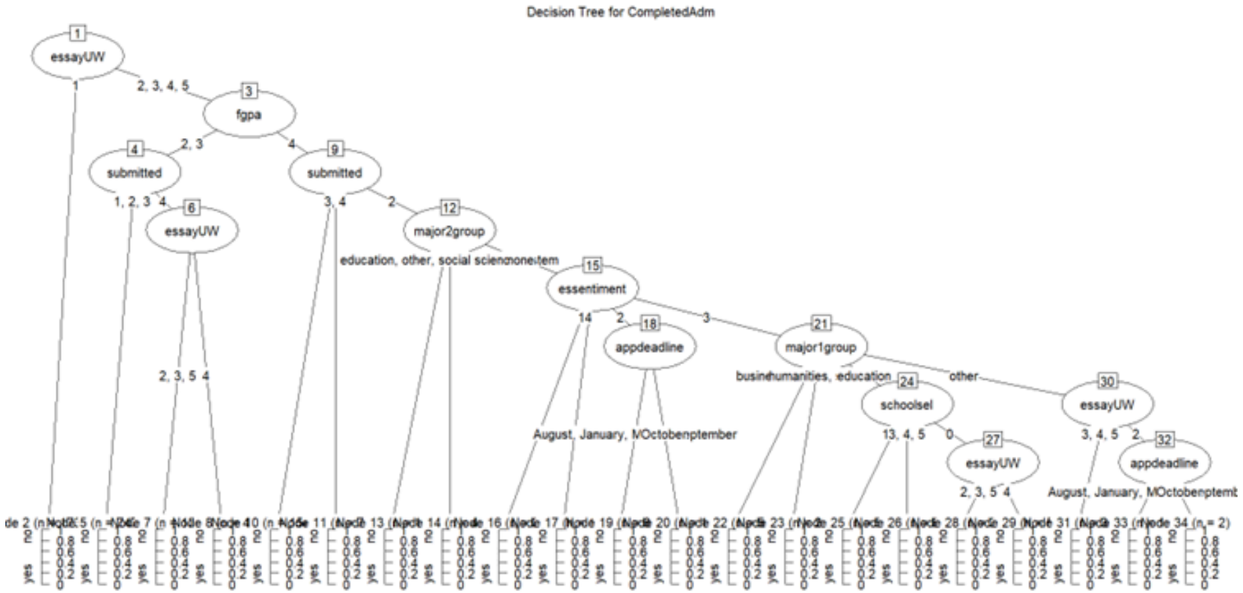
```
> error_cost_DT
     no yes
no    1   2
yes   4   1
```

Result:

- Retrained model with a more complex structure (20 nodes).

Reference / prediction	0	1
0	24	1215
1	42	4968

- **Accuracy**: 0.7988
- **Sensitivity:** 0.363636
- **Specificity:** 0.803493



5

Artificial Neural Network

Artificial Neural Network Models

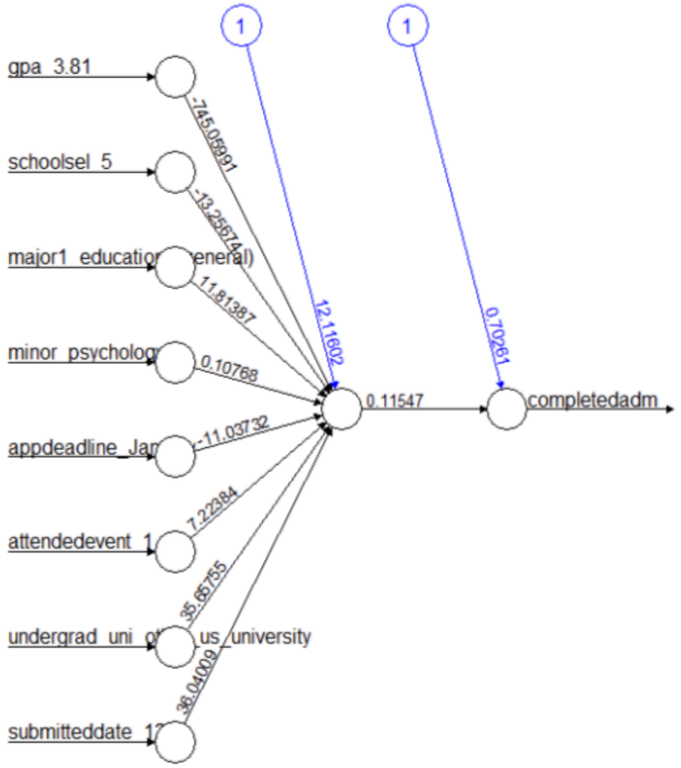
Ideal Candidate Variables

- **GPA (3.81):** Strong academic preparedness.
- **School Selectivity (5):** Motivation from attending a highly selective institution.
- **Major/Minor:** Education (primary) and Psychology (minor) for program alignment.
- **Application Deadline (Jan):** Reflects time management.
- **Event Attendance (1):** Shows interest and engagement.
- **Undergrad (US):** Adds background diversity.
- **Submission Date (120):** Indicates punctuality and commitment.

Objective of ANN Models:

- Identify patterns in ideal candidates who complete the program vs. those likely to withdraw.
- Focus on key variables that predict program completion, avoiding noise and imbalanced data.

Artificial Neural Network Models 1



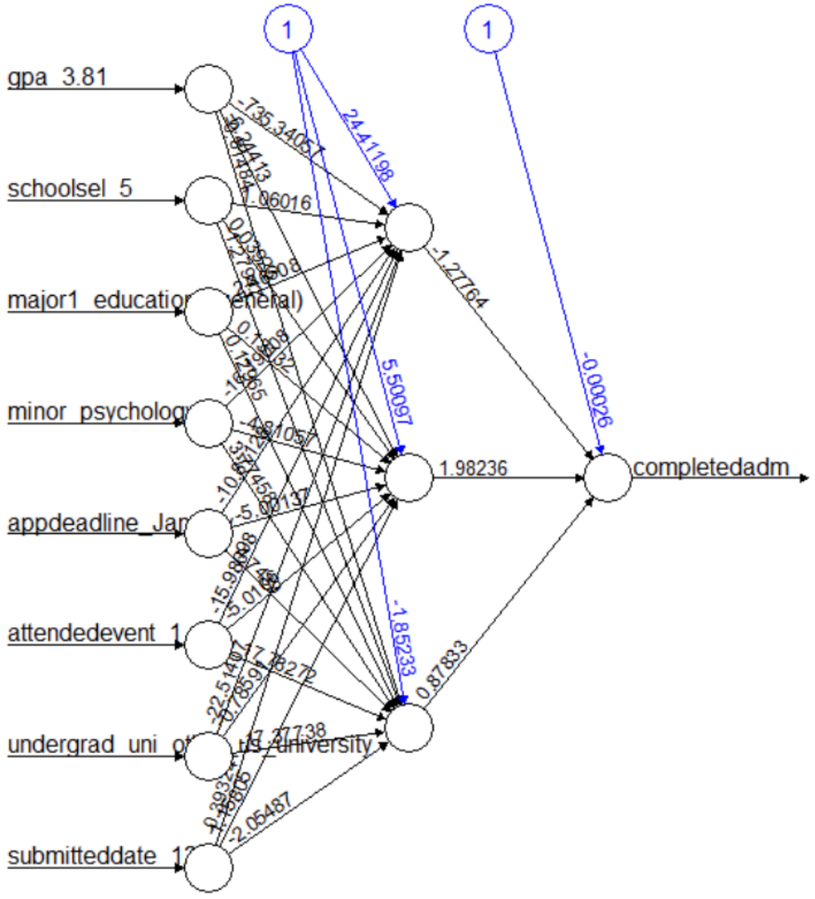
Error: 1728.023031 Steps: 457203

ANN Model 1: Confusion Matrix and Performance Statistics

Reference / prediction	0	1
0	1862	7511
1	0	0

- **Accuracy**: 0.1987 (95% CI: 0.1906, 0.2069)
- **No Information Rate**: 0.8013
- **Kappa**: 0
- **McNemar's Test P-Value**: < 2e-16
- **Sensitivity**: 1.0000 (ability to identify applicants likely to withdraw)
- **Specificity**: 0.0000 (ability to identify applicants likely to complete)
- **Positive Predictive Value**: 0.1987
- **Negative Predictive Value**: NaN
- **Balanced Accuracy**: 0.5000

Artificial Neural Network Models 2



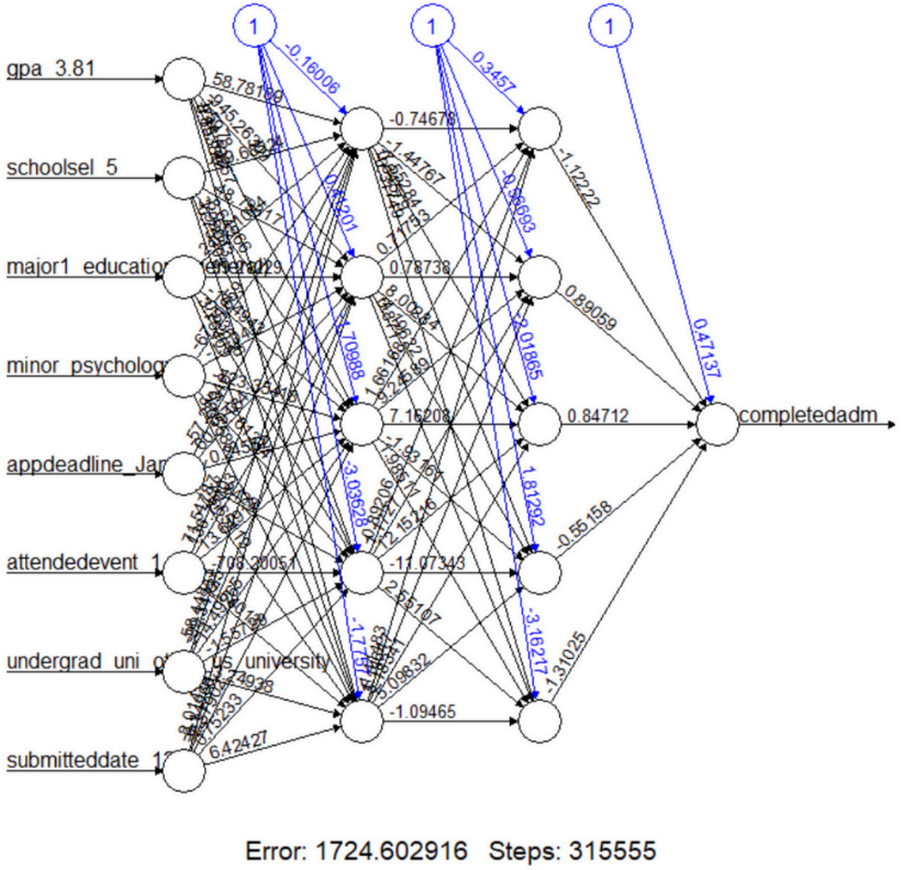
Error: 1725.505884 Steps: 669752

ANN Model 1: Confusion Matrix and Performance Statistics

Reference / prediction	0	1
0	1862	7511
1	0	0

- **Accuracy**: 0.1987 (95% CI: 0.1906, 0.2069)
- **No Information Rate**: 0.8013
- **Kappa**: 0
- **McNemar's Test P-Value**: < 2e-16
- **Sensitivity**: 1.0000 (ability to identify applicants likely to withdraw)
- **Specificity**: 0.0000 (ability to identify applicants likely to complete)
- **Positive Predictive Value**: 0.1987
- **Negative Predictive Value**: NaN
- **Balanced Accuracy**: 0.5000

Artificial Neural Network Models 3



ANN Model 1: Confusion Matrix and Performance Statistics

Reference / prediction	0	1
0	1862	7511
1	0	0

- **Accuracy**: 0.1987 (95% CI: 0.1906, 0.2069)
- **No Information Rate**: 0.8013
- **Kappa**: 0
- **McNemar's Test P-Value**: < 2e-16
- **Sensitivity**: 1.0000 (ability to identify applicants likely to withdraw)
- **Specificity**: 0.0000 (ability to identify applicants likely to complete)
- **Positive Predictive Value**: 0.1987
- **Negative Predictive Value**: NaN
- **Balanced Accuracy**: 0.5000

Artificial Neural Network Models

Model	Hidden Layers/Nodes	Accuracy	Sensitivity	Specificity	Observations
Model 1	No Hidden Layer	0.1987	1	0	Identifies all candidates as “withdrawals”; low predictive value for completed.
Model 2	1 Layer, 3 Hidden Nodes	0.1987	1	0	Three nodes did not improve performance; specificity remained zero.
Model 3	2 Layers, 5 Nodes Each	0.1987	1	0	Additional layers and nodes showed no improvement; specificity remained zero.

Observations:
All models achieved high sensitivity (1.0), accurately identifying “Completed” cases.
Low specificity (0.0) in all models indicates failure to identify withdrawals, revealing model bias towards predicting "Completed."

Artificial Neural Network Models

Model Insights:

- **Class Imbalance:** Models overfit the “Withdraw” class, showing high sensitivity but poor specificity.
- **Low Predictive Accuracy:** Accuracy consistently shows as 19.87% (0.1987).
- This profile represents a highly committed candidate who is more likely to **withdraw the application process**.

Recommendations:

1. Test alternative models (e.g., decision trees, Naive Bayes) for improved performance.
2. Improved Techniques: Apply features to better capture “completed” risk.

6

Support Vector Machine

Model 1: Mixed Features with Categorical and Numeric Predictors

- Predictors include mix of applicant information, essay characteristics, application deadlines, and event attendance status
  - *gpa, stem, schoolsel, major1group, major2group, minorgroup, essayuniquewords, essaysentiment, appdeadline, submitteddate, attendedevent, completedadm*
- Used Linear (‘vanilladot’) & RBF Kernel (‘rbfdot’)

Summary of Results:

- Accuracy of 80.17% but undefined positive predictive value
- Failed to predict any withdrawn applications

MODEL 1

Reference/Prediction	0	1
0	0	0
1	1239	5010

Accuracy: 0.8017
Sensitivity: 0.0000
Specificity: 1.0000
Positive Predictive Value: NaN
Negative Predictive Value: 0.8017

Model 2 and 3: Numeric Predictors

Predictors:

- Model 2 numeric predictors exclude <u>applicant's major</u>
  - *gpa, stem, schoolsel, essay1length, essay2length, essay3length, essayuniquewords, essaysentiment, signupdate, starteddate, appdeadline, submitteddate, attendedevent, and completedadm*

- Model 3 has the same predictors but <u>excludes 'essaylength' variables</u>

Model 2

Reference/Prediction	0	1
0	6	0
1	1233	5010

- **Accuracy**: 0.8027 (95% CI: 0.7926, 0.8125)
- **No Information Rate**: 0.8017
- **Kappa**: 0.0077
- **McNemar's Test P-Value**: < 2e-16
- **Sensitivity**: 0.00484
- **Specificity**: 1.00000
- **Positive Predictive Value**: 1.00000
- **Negative Predictive Value**: 0.80249
- **Balanced Accuracy**: 0.50242

Model 3

Reference/Prediction	0	1
0	5	0
1	1234	5010

- **Accuracy**: 0.8025 (95% CI: 0.7924, 0.8123)
- **No Information Rate**: 0.8017
- **Kappa**: 0.0065
- **McNemar's Test P-Value**: < 2e-16
- **Sensitivity**: 0.00403
- **Specificity**: 1.00000
- **Positive Predictive Value**: 1.00000
- **Negative Predictive Value**: 0.80237
- **Balanced Accuracy**: 0.50201

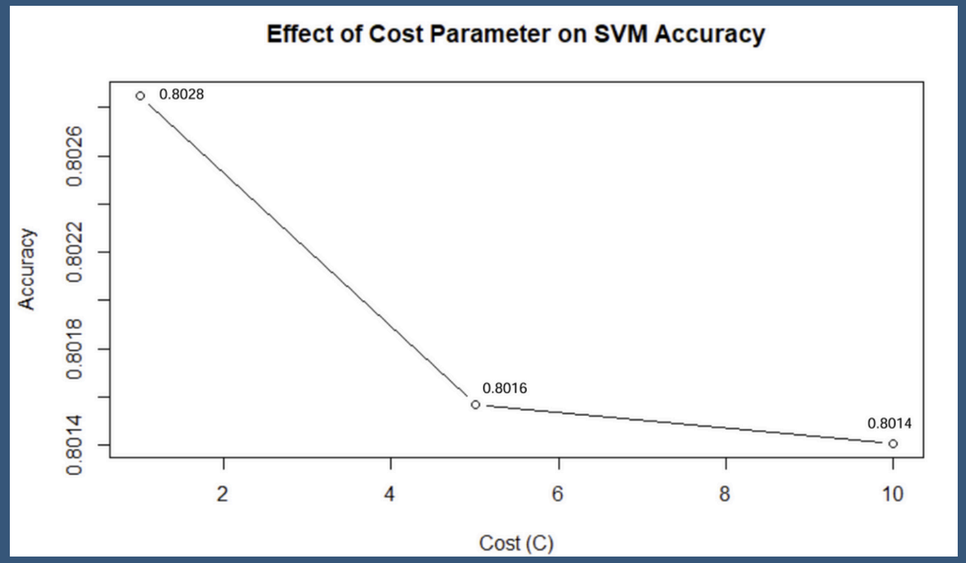
RBF Kernel Results

Summary of Results:

- Linear Kernel results similar to Model 1
- Positive predictive value is no longer undefined
- Improved accuracy (from 80.17%) and increased sensitivity (from 0)

Final Model

- Refined numeric predictors - same predictors but excluding essay lengths, sign up date, and start date
  - *gpa, stem, schoolsel, essayuniquewords, essaysentiment, appdeadline, submitteddate, attendedevent, and completedadm*



Cost Parameters vs Accuracy
C=~1 : 0.8028
C=5 : 0.8016
C = 10 : 0.8014

MODEL 4

Reference/Prediction	0	1
0	7	0
1	1232	5010

Accuracy: 0.8028
Sensitivity: 0.00565
Specificity: 1.0000
Positive Predictive Value: 1.0000
Negative Predictive Value: 0.80263

RBF Kernel Results

Summary of Results:

- Improved results for all values!
- Slightly outperforms previous models
  - by excluding less relevant predictors
- Linear kernel remains the same as Model 1

7

CONCLUSION

Best models summary

MODEL	ACCURACY	SENSITIVITY	SPECIFICITY	KAPPA	BALANCED ACURACY
NB MODEL 4	0.7969	0.06457	0.97804	0.0625	0.52131
DT MODEL 2	0.7988	0.363636	0.803493	0.0171	0.583565
ANN MODEL 3	0.1987	1	0	0	0.5
SVM MODEL 4	0.8028	0.00565	1	0.4195	0.50282

THANK YOU !